Package 'semicontMANOVA'

January 10, 2024

Title Multivariate ANalysis of VAriance with Ridge Regularization for Semicontinuous High-Dimensional Data

LazyLoad yes

Version 0.1-8

Depends R (>= 2.15.1)

Imports matrixcalc, mvtnorm

Date 2024-01-06

Description Implements Multivariate ANalysis Of VAriance (MANOVA) parameters' inference and test with regularization for semicontinuous high-dimensional data. The method can be applied also in presence of low-dimensional data. The p-value can be obtained through asymptotic distribution or using a permutation procedure. The package gives also the possibility to simulate this type of data. Method is described in Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni (2024) <arXiv:2401.04036>.

License GPL-2

NeedsCompilation no

Author Elena Sabbioni [aut, cre] (<https://orcid.org/0000-0002-8099-1216>), Claudio Agostinelli [aut] (<https://orcid.org/0000-0001-6702-4312>), Alessio Farcomeni [aut] (<https://orcid.org/0000-0002-7104-5826>)

Maintainer Elena Sabbioni <elena.sabbioni@polito.it>

Repository CRAN

Date/Publication 2024-01-10 19:30:02 UTC

R topics documented:

scMANOVA *Multivariate ANalysis Of VAriance Inference and Test with Ridge Regularization for Semicontinuous High-Dimensional Data*

Description

scMANOVA performs Multivariate ANalysis Of VAriance (MANOVA) inference and test with ridge regularization in presence of semicontinuous high-dimensional data. The test is based on a Likelihood Ratio Test statistic and the p-value can be computed using either asymptotic distribution (p.value.perm = FALSE) or via permutation procedure (p.value.perm = TRUE). There is the possibility to provide as input the regularization parameters or to choose them through an optimization procedure.

Usage

```
scMANOVA(x, n, lambda = NULL, lambda0 = NULL, lambda.step = 0.1,
  ident = FALSE, tol = 1e-08, penalty = function(n, p) log(n),
  B = 500, p.value.perm = FALSE, fixed.lambda = FALSE, ...)
```

Arguments

x	data.frame or matrix of data with units on the rows and variables on the columns
n	vector. The length corresponds to the number of groups, the elements to the number of observations in each group
lambda	NULL, a scalar or a vector of length 2. Ridge regularization parameter. The optimal value of lambda is searched in the interval [0,100] if NULL, and in the specified interval when it is a vector of length 2, otherwise it is used as the optimal value
lambda0	NULL, a scalar or a vector of length 2. Ridge regularization parameter under null hypothesis. The optimal value of lambda0 is searched in the interval [0,100] if NULL, and in the specified interval when it is a vector of length 2, otherwise it is used as the optimal value
lambda.step	scalar. Step size used in the optimization procedure to find the smallest value of lambda (and lambda0) that makes the covariance matrices, under the alternative and under the null hypotheses, non singular
ident	logical. If TRUE, lambda times the identity matrix is added to the raw estimated covariance matrix, if FALSE the diagonal values of the raw estimated covariance matrix are used instead
tol	scalar. Used in the optimization procedure to find the smallest value of lambda (and lambda0) that makes the covariance matrices, under the alternative and under the null, non singular
penalty	function with two arguments: sample size (n) and number of variables (p) used as penalty function in the definition of the Information Criterion to select the optimal values for lambda and lambda0

B	scalar. Number of permutations to run in the permutation test
p.value.perm	logical. If TRUE a p-value from a permutation test is evaluated, otherwise an asymptotic value is reported
fixed.lambda	logical. If TRUE the optimal values for lambda and lambda0 are evaluated just once for the observed dataset and kept fixed during the permutation test, otherwise, optimal values are evaluated for each permuted datsets
...	further parameters passed to function scMANOVApermTest

Value

An object of class scMANOVA which is a list with the following components

pi	matrix. Estimated proportion of missing values for each group
mu	matrix. Estimated mean vector for each group
sigmaRidge	matrix. Estimated covariance matrix with ridge regularization
sigma	matrix. Estimated covariance matrix by maximum likelihood
pi0	vector. Estimated proportion of missing values under the null hypothesis
mu0	vector. Estimated mean vector under the null hypothesis
sigma0Ridge	matrix. Estimated covariance matrix with ridge regularization under null hypothesis
sigma0	matrix. Estimated covariance matrix by maximum likelihood under null hypothesis
removed.vars	vector or NULL. columns removed in the continuous part of the log-likelihood dues to insufficient number of observations in each group
logLikPi	scalar. Log-likelihood for the discrete part of the model
logLik	scalar. Log-likelihood
logLikPi0	scalar. Log-likelihood for the discrete part of the model under the null hypothesis
logLik0	scalar. Log-likelihood under null hypothesis
statistic	scalar. Wilks statistics
lambda	scalar. Regularization parameter
lambda0	scalar. Regularization parameter under null hypothesis
df	scalar. Model degree of freedom
df0	scalar. Model degree of freedom under null hypothesis
aic	scalar. Information criteria
aic0	scalar. Information criteria under null hypothesis
p.value	scalar. p-value of the Wilks statistic

Author(s)

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni

References

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni (2024) A regularized MANOVA test for semicontinuous high-dimensional data. arXiv: http://arxiv.org/abs/2401.04036

See Also

scMANOVAestimation and scMANOVApermTest

Examples

```
set.seed(1234)
n <- c(5,5)
p <- 20
pmiss <- 0.1
x <- scMANOVAsimulation(n=n, p=p, pmiss=pmiss)
res.asy <- scMANOVA(x=x, n=n) # Asymptotic p.value
res.asy

  res.perm <- scMANOVA(x=x, n=n, p.value.perm=TRUE) # p-value by permutation test
  res.perm
```

`scMANOVAestimation` *Multivariate ANalysis Of VAriance Maximum Likelihood Estimation with Ridge Regularization for Semicontinuous High-Dimensional Data*

Description

`scMANOVAestimation` computes the regularized Multivariate ANalysis Of VAriance (MANOVA) maximum likelihood estimates for semicontinuous high-dimensional data. The estimation can be performed also for low-dimensional data. The regularization parameters are provided as input and the user can decide to perform the regularization adding the identity matrix to the raw estimated covariance matrix (default, `ident=TRUE`) or adding the diagonal values of the raw estimated covariance matrix (`ident=FALSE`).

Usage

```
scMANOVAestimation(x, n, lambda = NULL, lambda0 = NULL,
  ident = TRUE, posdef.check = TRUE, rm.vars = NA)
```

Arguments

`x`	`data.frame` or `matrix` of data with units on the rows and variables on the columns
`n`	`vector`. The length corresponds to the number of groups, the elements to the number of observations in each group
`lambda`	scalar. Ridge regularization parameter

lambda0	scalar. Ridge regularization parameter under null hypothesis
ident	logical. If TRUE, lambda times the identity matrix is added to the raw estimated covariance matrix, if FALSE the diagonal values of the raw estimated covariance matrix are used instead
posdef.check	logical. Check if the estimated covariance matrix is positive definite
rm.vars	NA, NULL or vector. If NA variables are removed from the analysis when they do not have enough observations to compute covariances. If NULL or a zero length vector all the variables are retained. If it is a vector, it indicates the position of the variables to remove, no further variables are removed

Value

An object of class scMANOVAestimation which is a list with the following components

pi	matrix. Estimated proportion of missing values for each group
mu	matrix. Estimated mean vector for each group
sigmaRidge	matrix. Estimated covariance matrix with ridge regularization
sigma	matrix. Estimated covariance matrix by maximum likelihood
pi0	vector. Estimated proportion of missing values under the null hypothesis
mu0	vector. Estimated mean vector under the null hypothesis
sigma0Ridge	matrix. Estimated covariance matrix with ridge regularization under null hypothesis
sigma0	matrix. Estimated covariance matrix by maximum likelihood under null hypothesis
removed.vars	vector or NULL. columns removed in the continuous part of the log-likelihood dues to insufficient number of observations in each group
logLikPi	scalar. Log-likelihood for the discrete part of the model
logLik	scalar. Log-likelihood
logLikPi0	scalar. Log-likelihood for the discrete part of the model under the null hypothesis
logLik0	scalar. Log-likelihood under null hypothesis

Author(s)

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni

References

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni (2024) A regularized MANOVA test for semicontinuous high-dimensional data. arXiv: http://arxiv.org/abs/2401.04036

See Also

scMANOVA and scMANOVApermTest

Examples

```
set.seed(1234)
n <- c(5,5)
p <- 20
pmiss <- 0.1
x <- scMANOVAsimulation(n=n, p=p, pmiss=pmiss)
res <- scMANOVAestimation(x=x, n=n, lambda=3.59, lambda0=3.13)
res
```

`scMANOVApermTest` *Multivariate ANalysis Of VAriance log-likelihood Test with Ridge Regularization for Semicontinuous High-Dimensional Data*

Description

`scMANOVApermTest` uses a permutation procedure to perform a test based on a Multivariate ANalysis Of VAriance(MANOVA) Likelihood Ratio test statistic with a ridge regularization. The statistic is developed for semicontinuous and high-dimensional data, but can be used also in low-dimensional scenarios.

Usage

```
scMANOVApermTest(x, n, lambda = NULL, lambda0 = NULL, lambda.step = 0.1,
  ident = FALSE, tol = 1e-08, penalty = function(n, p) log(n), B = 500,
  parallel = c("no", "multicore", "snow"), ncpus = 1L, cl = NULL,
  only.pvalue = TRUE, rm.vars = NA, ...)
```

Arguments

`x`	`data.frame` or `matrix` of data with units on the rows and variables on the columns
`n`	`vector`. The length corresponds to the number of groups, the elements to the number of observations in each group
`lambda`	scalar or a `vector` of length 2. Ridge regularization parameter. The optimal value of `lambda` is searched in the specified interval when it is a vector of length 2, otherwise it is used as the optimal value
`lambda0`	`NULL`, a scalar or a `vector` of length 2. Ridge regularization parameter under null hypothesis. The optimal value of `lambda0` is searched in the specified interval when it is a vector of length 2, otherwise it is used as the optimal value
`lambda.step`	scalar. Step size used in the optimization procedure to find the smallest value of `lambda` (and `lambda0`) that makes the covariance matrices, under the alternative and under the null hypothesis, non singular
`ident`	`logical`. If `TRUE`, `lambda` times the identity matrix is added to the raw estimated covariance matrix, if `FALSE` the diagonal values of the raw estimated covariance matrix are used instead

tol	scalar. Used in the optimization procedure to find the smallest value of lambda (and lambda0) that makes the covariance matrices, under the alternative and under the null hypothesis, non singular
penalty	function with two arguments: sample size (n) and number of variables (p) used as penalty function in the definition of the Information Criterion to select the optimal values for lambda and lambda0
B	scalar. Number of permutations to run in the permutation test
parallel	The type of parallel operation to be used (if any)
ncpus	integer. Number of processes to be used in parallel operation: typically one would chose this to the number of available CPUs.
cl	An optional parallel or snow cluster to use if parallel = "snow". If not supplied, a cluster on the local machine is created for the duration of the call
only.pvalue	logical. If TRUE only the p-value is returned
rm.vars	vector. It indicates the position of the variables to remove
...	Further parameters passed to parallel::mclapply in case of parallel="multicore"

Value

If only.pvalue=TRUE (default) a scalar which is the p-value of the Wilks statistic obtain by a permutation procedure, otherwise an object of class htest

Author(s)

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni

References

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni (2024) A regularized MANOVA test for semicontinuous high-dimensional data. arXiv: http://arxiv.org/abs/2401.04036

See Also

scMANOVA and scMANOVAestimation

Examples

```
set.seed(1234)
n <- c(5,5)
p <- 20
pmiss <- 0.1
x <- scMANOVAsimulation(n=n, p=p, pmiss=pmiss)
res <- scMANOVApermTest(x=x, n=n, lambda=3.59, lambda0=3.13,
  only.pvalue=FALSE)
res
```

scMANOVAsimulation *Simulation of datasets for a semicontinuous scenarios*

Description

Simulation of dataset of semicontinuous data coming from different groups, with specific marginal probabilities of a missing value, specific mean vectors and common covariance matrix.

Usage

```
scMANOVAsimulation(n, p, pmiss = 0, rho = 0, mu = NULL,
  sigma = NULL, only.data = TRUE)
```

Arguments

n	vector. The length corresponds to the number of groups, the elements to the number of observations in each group
p	scalar. Number of variables (columns)
pmiss	scalar or vector. Proportion of missingness in each group. If it is a scalar the same proportion is used in each group
rho	scalar. If sigma=NULL then sigma is set as a covariance matrix with covariance rho equal in every entries that is not on the main diagonal of sigma, and variance equal to 1
mu	NULL or vector or matrix. If NULL the mean of each group is set zero for all the variables, if vector the different groups have the same mean. If matrix each row corresponds to the mean vector of the corresponding group
sigma	NULL or matrix. If matrix it is a covariance matrix. If NULL the value of rho is used to build the covariance matrix
only.data	logical. If TRUE only the simulated data are reported

Value

If only.data=TRUE an object of class matrix is reported otherwise a list with the following components

x	matrix. The simulated dataset
y	matrix. A matrix with zero when the corresponding entry in x is zero and one otherwise
original	matrix. The simulated dataset without missing values
mu	matrix. Mean vectors, on for each group
sigma	matrix. Covariance matric
n	As in input
p	As in input
pmiss	vector. Proportion of missingness in each group

Author(s)

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni

References

Elena Sabbioni, Claudio Agostinelli and Alessio Farcomeni (2024) A regularized MANOVA test for semicontinuous high-dimensional data. arXiv: http://arxiv.org/abs/2401.04036

See Also

scMANOVAestimation and scMANOVApermTest

Examples

```
set.seed(1234)
n <- c(5,5)
p <- 20
pmiss <- 0.1
x <- scMANOVAsimulation(n=n, p=p, pmiss=pmiss)
```

Index